895562

ABSTRACT

Disclosed herein is a method for automatically filtering a corpus of documents containing textual and non-textual information of a natural language. According to the method, through a first dividing step (101), the document corpus is divided into appropriate portions. At a following determining step (105), for each portion of the document corpus, there is determined a regularity value (V_R) measuring the conformity of the portion with respect to character sequences probabilities predetermined for the language considered. At a comparing step (107), each regularity value (V_R) is then compared with a threshold value (V_T) to decide whether the conformity is sufficient. Finally, at a rejecting step (111), any portion of the document corpus whose conformity is not sufficient is rejected and removed from the corpus. An apparatus for carrying out such a method is also disclosed.

(FIGURE 1)